

NOTES 01: DATA

Stat 120 | Fall 2025

Prof Amanda Luby

1 The Structure of Data

Cases

Variables

2 Types of Variables

Quantitative

Categorical

Explanatory

Response

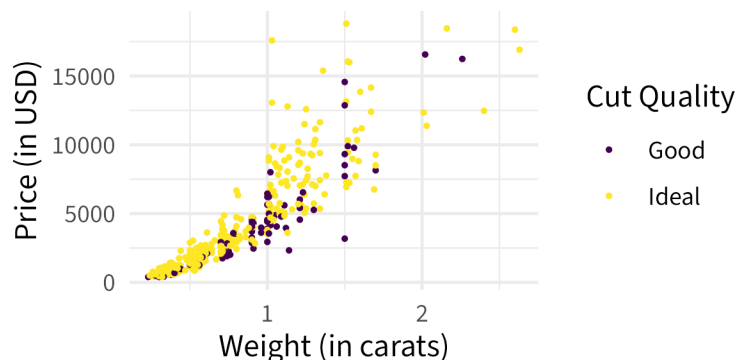
3 Examples

Label the *cases* and *variables*. For each variable, state whether it is *categorical* or *quantitative*. Indicate if there's clear *response* or *explanatory* variables

3.1 Penguins

```
# A tibble: 344 x 8
  species island bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
  <fct>   <fct>         <dbl>         <dbl>         <int>         <int>
1 Adelia Torgersen      39.1           18.7           181          3750
2 Adelia Torgersen      39.5           17.4           186          3800
3 Adelia Torgersen      40.3           18            195          3250
4 Adelia Torgersen      NA            NA             NA            NA
5 Adelia Torgersen      36.7           19.3           193          3450
6 Adelia Torgersen      39.3           20.6           190          3650
7 Adelia Torgersen      38.9           17.8           181          3625
8 Adelia Torgersen      39.2           19.6           195          4675
9 Adelia Torgersen      34.1           18.1           193          3475
10 Adelia Torgersen      42            20.2           190          4250
# i 334 more rows
# i 2 more variables: sex <fct>, year <int>
```

3.2 Price of diamonds by carat and cut quality



3.3 Is there a "Sprinting Gene"?

A gene called ACTN3 encodes a protein which functions in fast-twitch muscles. Some people have a variant of this gene that cannot yield this protein. To address the question of whether this gene is associated with sprinting ability, geneticists tested people from three different groups: world-class sprinters, world-class marathon runners, and a control group of non-athletes. In the same test, 6% of the sprinters had the gene variant, compared with 18% of non-athletes and 24% of the marathon runners.

Sketch out a possible dataset, and then answer the questions.